EMPIRICAL BAYES ESTIMATION IN HETEROGENEOUS MATCHED BINARY SAMPLES WITH SYSTEMATIC AGING EFFECTS*

Bruce Levin

Columbia University

We discuss two empirical Bayes estimation problems for heterogeneous matched binary samples with systematic growth effects, in the applied study of recognized spontaneous abortion. The first problem is to estimate an assumed systematic component in the random growth curve, and sufficient conditions are provided for consistent estimation of governing structural parameters. The second problem is to estimate future risk based on past outcomes, and for this we extend Robbins' general empirical Bayes estimator for binomial variables to the case of a sum of conditionally independent, non-identically distributed binary variables.

1. Introduction.

There is an interesting application of empirical Bayes estimation in the study of recognized spontaneous abortion (miscarriage). We specify a mathematical model for binary outcome data that incorporates several factors identified by epidemiologists as necessary for a realistic analysis of obstetric sequences. The factors are heterogeneity of risk, systematic effects of maternal age and gravidity, selective fertility, and differential pregnancy

*This research was supported by NIH Contract 1-R01-HD-15909.

AMS 1980 subject classifications: Primary 62C12, 62F10, 62G05; Secondary 62P10, 60J20

Key words and phrases. General empirical Bayes estimation, heterogeneous and non-identically distributed binary variables, random growth curves, spontaneous abortion.

spacing. Two estimation problems are considered in the context of a cross-sectional survey: (I) estimation of the systematic effects of maternal age and gravidity, and (II) estimation of future risk based on past obstetric history. In the first problem, we identify sufficient conditions under which one obtains consistent estimates of structural age and gravidity effects that are free of the biases often encountered in marginal analyses of rates due to the factors of risk heterogeneity, selective fertility, and differential pregnancy spacing. An important feature of the model is that it allows the individual contributions of age and gravidity to be simultaneously assessed, thus addressing a controversy in the epidemiologic literature of the past decade. For the second estimation problem, we require an extension of Robbins' general empirical Bayes estimator for binomial variables to the case of a sum of conditionally independent, non-identically distributed binary variables.

An analysis of serially correlated binary variables has recently been given by Stiratelli, Laird, and Ware (1984) for longitudinal studies, extending the related work in normal theory of Laird and Ware (1982) and Hui and Berger (1983). The papers of Koch et al. (1977) and Korn and Whittemore (1979) discuss alternative analyses for repeated categorical observations. Our analysis differs from these authors' in the following respects: (i) we consider only the simplest case of a single random effect parameter for risk, with no random heterogeneity in the structural growth curves (all such heterogeneity is treated as systematic); (ii) we do not restrict ourselves to parametric families of priors, the empirical Bayes estimators being "general" in the sense of Robbins (1980, 1983) (although we consider only the simpler problem of estimating posterior odds as opposed to probabilities); (iii) the analysis is applicable to both longitudinal and survey designs in dynamically stable populations where explicit adjustments are necessary for the kinds of biasing factors considered herein. James (1969) has discussed a test for birth order effects in the presence of selective fertility which is close in spirit to the present paper.

In section 2 we briefly review some substantive findings that characterize the spontaneous abortion process among gravid women. The

discussion motivates the model assumptions specified in section 3. In section 4 we provide a consistent estimator for the structural growth parameters, and the extension of the general empirical Bayes estimator of future risk given observed obstetric history.

It is a pleasure to acknowledge Professor Robbins for his development of empirical Bayes theory, the elegance of which still shines through the details of this application. Robbins' interest in empirical Bayes methods for binary variables may be traced at least as far back as Robbins (1956), and as recently as his work on general and linear e.B. estimation for binomials, discussed in Cressie (1982), and Robbins (1983).

2. Substantive background from the epidemiology of spontaneous abortion.

The relations of maternal age, gravidity (birth order), and obstetric history to the risk of spontaneous abortion have proven difficult to separate. When marginal age-specific rates of spontaneous abortion are studied (as a proportion of spontaneous abortions plus term births), the risk is observed to increase at late maternal ages, beginning in the mid-thirties. It increases with gravidity, and also among women who have experienced one or more previous miscarriages ("recurrence risk"). These three observations are not unrelated. For example, maternal age and gravidity are positively correlated, and total gravidity and rate of spontaneous abortions may be associated in circumstances where the decision to become pregnant again is related to the outcomes of previous pregnancies. The apparent tendency toward recurrence is usually interpreted as an indication of underlying heterogeneity between women in their risk for spontaneous abortion (Warburton and Fraser, 1964; Naylor, 1974; James, 1974; Leridon, 1976; reviewed in Wilcox and Gladen, 1982). In this view, recurrence risk is a selection effect, where women who have experienced one or more miscarriages overrepresent women with a high risk of spontaneous abortion on any pregnancy, relative to the population as a whole. An alternative explanation -- that one spontaneous abortion is causally related to a later spontaneous abortion ("state dependence") -- has not been examined carefully,

although several authors point to this possibility as well (Naylor and Warburton, 1978; Wilcox and Gladen, 1982). The state dependence model will not be pursued here.

Previous discussions of risk heterogeneity have usually (James, 1974; Leridon, 1976; Wilcox and Gladen, 1982), but not always (Warburton and Fraser, 1964), focused on whether or not heterogeneity could have produced the maternal age and gravidity effects observed in the marginal rates of abortion. Besides heterogeneity, and the possibility of recall bias and evasive answer bias, several other factors have been identified that might give rise to spurious marginal associations between the risk of abortion and the characteristics of age, gravidity, and obstetric history. Much debate has arisen over the relative importance of age versus gravidity because of failures to account for one or more of these biasing factors. Leridon (1976) and Wilcox and Gladen (1982) provide the most comprehensive reviews of the issues that have been raised. Two factors in particular, selective fertility (wherein continuation of reproduction depends on the outcome of previous pregnancies), and differential spacing between pregnancies (on average shorter after miscarriage than after term birth) are cited as major sources of spurious associations. Our focus here is <u>not</u> on the ways in which a marginal analysis of pregnancy outcomes may be misleading. Instead, we provide an analytic framework for studying the influences of age and gravidity on abortion risk <u>within</u> women, and for estimating the future risk of spontaneous abortion for each woman interviewed.

3. <u>Model specification</u>.

Let Π denote a population of gravid women whose first pregnancy occurred since calendar date t_0. In a cross-sectional "door-to-door" survey conducted around date $t_1 > t_0$, information is gathered on respondents' <u>current gravidity</u> I, and for those women with $I \geq 1$, the age A_i and binary outcome X_i on each pregnancy $i=1,\ldots,I$ is obtained. Let $X^{(0)} = A^{(0)} = \phi$, and for $i \geq 1$, let $X^{(i)} = (X_1,\ldots,X_i)$ and $A^{(i)} = (A_1,\ldots,A_i)$. Each women surveyed in Π contributes the datum $(I,A^{(I)},X^{(I)})$.

Below we consider a class of coin-tossing models for the outcome process in which the outcomes are assumed to be conditionally independent within women given age, gravidity, and a random intensity parameter. The substantive considerations in section 2 lead us to make the following specific assumptions.

A1. (Heterogeneous risk) There exists an intensity parameter λ for each woman describing the log-odds on the event $[X_1 = 1]$, assuming that a first pregnancy were to occur at reference age A_0 which is fixed at, say, 20 years of age. The distribution of λ in Π, say $dH(\lambda)$ on $[-\infty,\infty]$, is assumed to satisfy

$$\int_{[\lambda<\infty]} e^{\lambda}\, dH(\lambda) < \infty ,$$

but is otherwise arbitrary and unknown.

A2. (Selective fertility) Each woman is assumed to follow a stopping rule N during her reproductive life indicating desired family size. We assume stopping may depend on the outcome sequence $X_1,\ldots,$ age $A_1,\ldots,$ and also on other factors that are completely independent of outcomes (such as sex of livebirths, accidental conception, or early infertility). Across women in Π, N varies as a random variable **N** defined on the class **C** of stopping rules just described. The distribution of **N** is unknown, but is assumed independent of λ,

$$dP_{\lambda}[\mathbf{N} = N] = dP[\mathbf{N} = N].$$

Below we write marginal conditional probability measures with the usual $P[\cdot\,|\,\cdot]$ notation, and we use subscript notation when conditioning also on given values of the intensity parameter and stopping rule, as in $P_{\lambda,N}[\cdot\,|\,\cdot]$.

A3$_N$. (Conditional independence of outcomes given age) Given λ and N, and the occurrence of an i-th pregnancy, the outcome X_i depends only on gravidity i and age A_i, but not on earlier outcomes or ages. That is,

$$P_{\lambda,N}[X_i = 1 \mid I \geq i,\ A^{(i)},\ X^{(i-1)}] = p_\lambda(i,A_i) = 1 - q_\lambda(i,A_i)$$

for some risk functions $p_\lambda(i,A_i)$ $(i=1,2,\ldots)$.

A4. (Structural age effect) There are structural growth functions $\phi_\beta(i,A)$, defined for $i \geq 1$ and $A > 0$, that depend for each i on a fixed number of unknown population parameters $\beta = \beta(i)$, such that $\phi_{\beta(1)}(1,A_0) = 0$, and for any A and λ,

$$\log\left[p_\lambda(i,A) \,/\, q_\lambda(i,A)\right] = \lambda + \phi_{\beta(i)}(i,A) .$$

Although ϕ depends on A_0, we suppress this from the notation. For simplicity here we assume no structural heterogeneity in the growth curves as a function of covariates. Some practical choices of ϕ_β include quadratic functions of $A-A_0$ or of $\log(A/A_0)$, e.g.

$$\phi_{\beta(1)}(1,A) = b_{11}\log(A / A_0) + b_{12}\log^2(A / A_0)$$

$$\phi_{\beta(i)}(i,A) = a_i + b_{i1}\log(A / A_0) + b_{i2}\log^2(A / A_0)$$

$$\beta(1) = (b_{11}, b_{12}) , \quad \beta(i) = (a_i, b_{i1}, b_{i2}).$$

The parameters a_i represent a gravidity effect superposed on the structural growth curves parameterized by b_{ij}.

$A5_N$. (Conditional independence of censoring plus fertility with respect to intensity) For $i \geq 1$ and any $A^{(i-1)}$, $X^{(i-1)}$, N,

$$P_{\lambda,N}[I \geq i \mid I \geq i-1,\ A^{(i-1)},\ X^{(i-1)}] \text{ is independent of } \lambda.$$

Assumption $A5_N$ declares the obstetric sequence $(A^{(i-1)}, X^{(i-1)})$ sufficient for the determination of the proportion of women who may be observed to have at

least i pregnancies, given i-1 or more, and stopping rule N. In the absence of recency censoring, accidental conception, and early infertility, assumption $A5_N$ is trivially satisfied, since for any λ the proportion of women observed to have at least i pregnancies given i-1 or more is either zero or one, depending only on the stopping rule N and the observed outcomes $X^{(i-1)}$. When recency censoring, accidental conception, or early infertility do occur, the proportions may be non-degenerate, and may depend on previous outcomes, but under $A5_N$ these effects apply independently of λ.

$A6_N$. (Conditional independence of pregnancy spacing wrt intensity) For $i \geq 1$ and any $A^{(i-1)}$, $X^{(i-1)}$, N there is a conditional probability density describing pregnancy spacing, say

$$P_{\lambda,N}[dA_i \mid I \geq i, A^{(i-1)}, X^{(i-1)}] \quad \text{that is independent of} \quad \lambda.$$

Thus the observable sequence up to gravidity i-1 suffices to determine pregnancy interval $A_i - A_{i-1}$ given $I \geq i$ and N. Note that the inter-pregnancy interval may depend differentially on previous outcomes $X^{(i-1)}$.

Assumptions $A3_N$, $A5_N$, and $A6_N$ have been stated in stronger form than we actually require, it being sufficient for our purposes to state these assumptions marginally with respect to N:

A3. $P_\lambda[X_i = 1 \mid I \geq i, A^{(i)}, X^{(i-1)}] = p_\lambda(i, A_i)$,

A5. $P_\lambda[I \geq i \mid I \geq i-1, A^{(i-1)}, X^{(i-1)}]$ is independent of λ,

A6. $P_\lambda[dA_i \mid I \geq i, A^{(i-1)}, X^{(i-1)}]$ is independent of λ.

We expect the assumptions to hold in gravid populations that are dynamically stable, in which pregnancy behavior is dependent only on past events observable to the woman, but not otherwise on the unobservable intensity parameter λ. The

importance of A5 and A6 is that the distributions of continuation and pregnancy spacing may be estimated from the marginal observed data. We now show that assumptions A2, $A3_N$, $A5_N$, and $A6_N$ imply A3, A5, and A6.

The likelihood of observing the event $[I \geq i, A^{(i)}, X^{(i)}]$ given λ and $\mathbf{N} = N$ is

$$P_{\lambda,N}[I \geq i, dA^{(i)}, X^{(i)}] = \prod_{j=1}^{i} P_{\lambda,N}[I \geq j \mid I \geq j-1, A^{(j-1)}, X^{(j-1)}]$$

$$\cdot \prod_{j=1}^{i} P_{\lambda,N}[dA_j \mid I \geq j, A^{(j-1)}, X^{(j-1)}]$$

$$\cdot \prod_{j=1}^{i} P_{\lambda,N}[X_j \mid I \geq j, A^{(j)}, X^{(j-1)}]$$

if $X^{(i)}$ is realizable under N, and is zero if not (e.g. if we exclude accidental fertility and early infertility). The likelihood is of the form

$$P_{\lambda,N}[I \geq i, dA^{(i)}, X^{(i)}] = \prod_{j=1}^{i} p_\lambda(j,A_j)^{X_j} q_\lambda(j,A_j)^{1-X_j} \cdot f(i, A^{(i)}, X^{(i)}, N)$$

where f does not depend on λ, and equals zero if $X^{(i)}$ is not realizable under N. Thus

$$dP_\lambda[\mathbf{N} = N \mid I \geq i, A^{(i)}, X^{(i)}] = \frac{P_{\lambda,N}[I \geq i, dA^{(i)}, X^{(i)}]\, dP[\mathbf{N} = N]}{\int_C P_{\lambda,N}[I \geq i, dA^{(i)}, X^{(i)}]\, dP[\mathbf{N} = N]}$$

$$= \frac{f(i, A^{(i)}, X^{(i)}, N)\, dP[\mathbf{N} = N]}{\int_C f(i, A^{(i)}, X^{(i)}, N)\, dP[\mathbf{N} = N]}$$

is independent of λ. Consequently

$$P_\lambda[I \geq i+1 \mid I \geq i, A^{(i)}, X^{(i)}] = \int_C P_{\lambda,N}[I \geq i+1 \mid I \geq i, A^{(i)}, X^{(i)}]$$

$$dP_\lambda[N = N \mid I \geq i, A^{(i)}, X^{(i)}]$$

is independent of λ, which is A5. A similar argument shows that both

$dP_\lambda[\mathbf{N} = N \mid I \geq i, A^{(i)}, X^{(i-1)}]$ and $dP_\lambda[\mathbf{N} = N \mid I \geq i, A^{(i-1)}, X^{(i-1)}]$ are independent of λ, implying A3 and A6, respectively.

4. Likelihood functions and estimation of structural parameters and future risk.

Let $g \geq 2$ be a given integer. Given λ, the likelihood of observing the event $[I=g, A^{(g)}, X^{(g)}]$ is, under assumptions A1 - A6,

$$(1)\quad P_\lambda[I=g, A^{(g)}, X^{(g)}] = \mu_g(A^{(g)}, X^{(g)})\, \nu_g(A^{(g)}, X^{(g)}) \prod_{i=1}^{g} p_\lambda(i, A_i)^{X_i} q_\lambda(i, A_i)^{1-X_i}$$

where

$$\mu_g(A^{(g)}, X^{(g)}) = \prod_{i=1}^{g} P[I \geq i \mid I \geq i-1, A^{(i-1)}, X^{(i-1)}] \cdot P[I = g \mid I \geq g, A^{(g)}, X^{(g)}]$$

and

$$\nu_g(A^{(g)}, X^{(g)}) = \prod_{i=1}^{g} P[dA_i \mid I \geq i, A^{(i-1)}, X^{(i-1)}]$$

are independent of λ. Note that ν_g depends on $X^{(g)}$ only through $X^{(g-1)}$. The product in (1) depending on λ can be written as

$$\exp(\lambda S_g) \exp\Big(\sum_{i=1}^{g} \phi_{\beta(i)}(i, A_i) X_i\Big) \Big/ \prod_{i=1}^{g} \{1 + \exp(\lambda + \phi_{\beta(i)}(i, A_i))\}$$

where $S_g = \sum_1^g X_i$. Thus given λ, the likelihood of $[I=g, A^{(g)}, S_g]$ is

$$P_\lambda[I=g, A^{(g)}, S_g] = c_g(S_g, A^{(g)}) \exp(\lambda S_g) \Big/ \prod_{i=1}^{g} \{1 + \exp(\lambda + \phi_{\beta(i)}(i, A_i))\}$$

where

$$(2)\quad c_g(s, A^{(g)}) = \sum_{x^{(g)}} \mu_g(A^{(g)}, x^{(g)})\, \nu_g(A^{(g)}, x^{(g)}) \exp\Big(\sum_{i=1}^{g} \phi_{\beta(i)}(i, A_i) x_i\Big),$$

the sum being taken over $D^{(g)}(s) = \{\text{all binary vectors } x^{(g)} \text{ with } x_1 + \ldots + x_g = s\}$.

4.1 Conditional likelihood analysis for the structural parameters.

Now let $0 < s < g$. Writing $x_1 = s - (x_2 + \ldots + x_g)$, the conditional likelihood function for $X^{(g)}$ given λ, $I = g$, $A^{(g)}$, and $S_g = s$ is

$$P_\lambda[X^{(g)} \mid I=g, A^{(g)}, S_g=s] = \mu_g(A^{(g)},X^{(g)})\,\nu_g(A^{(g)},X^{(g)}) \cdot \exp\{-\Psi + \sum_{i=2}^{g}[\phi_{\beta(i)}(i,A_i) - \phi_{\beta(1)}(1,A_1)]\,x_i\} \tag{3}$$

where

$$\Psi = \log\Big(\sum_{x^{(g)}} \mu_g(A^{(g)},x^{(g)})\,\nu_g(A^{(g)},x^{(g)}) \cdot \exp\Big\{\sum_{i=2}^{g}[\phi_{\beta(i)}(i,A_i) - \phi_{\beta(1)}(1,A_1)]\,x_i\Big\}\Big).$$

Equation (3) is an exponential family of distributions over $D^{(g)}(s)$ with respect to the dominating measure placing mass $\mu_g(A^{(g)},x^{(g)})\,\nu_g(A^{(g)},x^{(g)})$ at $x^{(g)} \in D^{(g)}(s)$, with sufficient statistics $X_2,\ldots,X_g$, and natural parameters

$$\phi_{\beta(i)}(i,A_i) - \phi_{\beta(1)}(1,A_1) \text{ for } i=2,\ldots,g.$$

Note that (3) does not depend on λ so that consistent estimation of β is possible.

The essence of (3) can be easily grasped by considering the simplest case $g = 2$, $S_2 = 1$. In this case (3) implies that the probability ratio of outcome $X^{(2)} = (0,1)$ to $X^{(2)} = (1,0)$ will be a product of three factors:

$$\frac{P[X^{(2)}=(0,1) \mid I=2, A^{(2)}, S_2=1]}{P[X^{(2)}=(1,0) \mid I=2, A^{(2)}, S_2=1]} = \left[\frac{\mu_2(A^{(2)},(0,1))}{\mu_2(A^{(2)},(1,0))}\right]\left[\frac{\nu_2(A^{(2)},(0,1))}{\nu_2(A^{(2)},(1,0))}\right] \cdot \exp[\phi_{\beta(2)}(2,A_2) - \phi_{\beta(1)}(1,A_1)].$$

Even without any age/gravidity effect, $\phi \equiv 0$, the outcomes (0,1) and (1,0) will

not in general be equally likely, but will depend on selective fertility (reflected in the first factor) and pregnancy spacing (reflected in the second factor). For example, among women with two closely spaced pregnancies and $S_2 = 1$, we expect to find more (1,0) outcomes than (0,1) outcomes if differential pregnancy spacing occurs as described in section 2, so that the second factor is less than one. Likelihood (3) represents a generalization of these remarks to the case of g pregnancies with $A^{(g)}$ and S_g fixed, and enables one to adjust the estimate of the structural growth component in the third factor for these null expectations.

For data analysis we need to consider a few practical matters and simplifications.

(a) Obtaining estimates of the dominating measure for large g will typically be difficult, so that we may wish to utilize information only up to some maximum gravidity G for those women with $I \geq G$. The truncated likelihood function will then be based on events $[I \geq G, A^{(G)}, X^{(G)}]$ rather than on the complete observed sequence. The only modification required is that the final factor in $\mu_G(A^{(G)},X^{(G)})$ is omitted. We shall adopt a maximum G, the truncated μ, ν, and truncated likelihood function without special notation.

(b) Assuming linear dependence of ϕ on the parameters $\beta(i)$, the log-likelihood function based on n informative women (with $0 < S_g < g$) is

$$(4) \quad \ell(\beta) = \Big(\sum_{\alpha=1}^{n} \sum_{j=2}^{g(\alpha)} X_{\alpha j} k'_{\alpha j} \Big)\beta - \sum_{\alpha=1}^{n} \log \Big\{ \sum_{u} \mu_{g(\alpha)} \cdot \nu_{g(\alpha)} \cdot \exp \Big[\Big(\sum_{j=2}^{g(\alpha)} u_j k'_{\alpha j} \Big)\beta \Big] \Big\}$$

where for the α^{th} woman with observed gravidity I, $g(\alpha) = \min(I,G)$, the sum inside the logarithm extends over $u \in D^{g(\alpha)}(S_{g(\alpha)})$, $k'_{\alpha j}$ is a row-vector of carrier age functions, and β is a column vector of parameters. For the example following assumption A4, we have for $2 \leq j \leq g(\alpha) \leq G$

$$k'_{\alpha j} = [\, -\ell_1, -\ell_1^2, 0, 0, 0, \ldots, 1, \ell_j, \ell_j^2, \ldots, 0, 0, 0]$$

where $\ell_j = \log (A_j / A_0)$, $(j \geq 1)$, and

$$\beta' = [b_{11}, b_{12}, a_2, b_{21}, b_{22}, \ldots, a_j, b_{j1}, b_{j2}, \ldots, a_G, b_{G1}, b_{G2}] \,.$$

Maximizing (4) is now a standard calculation, and if the components of the conditional information matrix $J(\beta) = \left(-\frac{\partial^2 \ell(\beta)}{\partial\beta_i \partial\beta_j}\right)$ grow unbounded as $n \to \infty$, the conditional maximum likelihood estimate $\hat{\beta}$ will be consistent with estimated asymptotic covariance matrix $J^{-1}(\hat{\beta})$.

(c) The dominating measure in (3) can be expressed in relative terms by dividing numerators and denominators by the mass at reference point $x_*^{(g)} = (1,\ldots,1, 0,\ldots,0) \in D^{(g)}(s)$. The measure then becomes $\mu_g^* \cdot \nu_g^*$, where

$$\mu_g^*(A^{(g)},x^{(g)}) = \frac{\mu_g(A^{(g)},x^{(g)})}{\mu_g(A^{(g)},x_*^{(g)})} = \prod_{i=2}^{g} \frac{P[I \geq i \mid I \geq i-1, A^{(i-1)}, x^{(i-1)}]}{P[I \geq i \mid I \geq i-1, A^{(i-1)}, x_*^{(i-1)}]}$$

$$\cdot \frac{P[I = g \mid I \geq g, A^{(g)}, x^{(g)}]}{P[I = g \mid I \geq g, A^{(g)}, x_*^{(g)}]}$$

and similarly

$$\nu_g^*(A^{(g)},x^{(g)}) = \frac{\nu_g(A^{(g)},x^{(g)})}{\nu_g(A^{(g)},x_*^{(g)})} = \prod_{i=2}^{g} \frac{P[dA_i \mid I \geq i, A^{(i-1)}, x^{(i-1)}]}{P[dA_i \mid I \geq i, A^{(i-1)}, x_*^{(i-1)}]}.$$

The measure ν_g^* simplifies considerably under a stationary Markov assumption such as

A7. For $i \geq 2$, $P[dA_i \mid I \geq i, A^{(i-1)}, x^{(i-1)}] = Q[dA_i \mid A_{i-1}, x_{i-1}]$ for some fixed density $Q[\cdot \mid A,x]$.

Under A7,

$$\nu_g^*(A^{(g)}, x^{(g)}) = \prod_{i=2}^{g} \frac{Q[dA_i \mid A_{i-1}, x_{i-1}]}{Q[dA_i \mid A_{i-1}, x_{*(i-1)}]}$$

$$= \prod_{i=2}^{s+1} \left\{\frac{Q[dA_i \mid A_{i-1}, 1]}{Q[dA_i \mid A_{i-1}, 0]}\right\}^{-1} \cdot \prod_{i=2}^{g} \left\{\frac{Q[dA_i \mid A_{i-1}, 1]}{Q[dA_i \mid A_{i-1}, 0]}\right\}^{x_{i-1}}$$

depending on $A^{(g)}$ only through the likelihood ratios $\frac{Q[\,\cdot \mid A, 1]}{Q[\,\cdot \mid A, 0]}$.

On substantive grounds, assumption A7 appears plausible, corresponding to an assumption that inter-pregnancy intervals (given that they are observed) depend only on the age and outcome at the previous pregnancy (see Leridon, 1976). A stationary Markov assumption for μ^* is not plausible, although a one-step Markov assumption in $A^{(i)}$ and $x^{(i)}$ may be tenable (see James, 1974), e.g.

A8. For $i \geq 2$,

$$P[I \geq i \mid I \geq i-1, A^{(i-1)}, x^{(i-1)}] = P[I \geq i \mid I \geq i-1, A_{i-1}, x_{i-1}].$$

(d) For continuous age models, obtaining μ_g^* involves estimation of binary regressions of continuation against age and previous outcomes, and obtaining ν_g^* involves estimation of smoothed density ratios. To avoid problems of density estimation, Bayes' theorem may be useful in that the desired quantities are related to the marginal odds on abortion given $A^{(g)}$ and given $A^{(g-1)}$. For example,

$$\frac{Q[dA_2 \mid A_1, X_1 = 1]}{Q[dA_2 \mid A_1, X_1 = 0]} = \frac{P[X_1 = 1 \mid A^{(2)}]}{P[X_1 = 0 \mid A^{(2)}]} \Big/ \frac{P[X_1 = 1 \mid A_1]}{P[X_1 = 0 \mid A_1]}.$$

Note that the marginal logistic regressions of outcome given age are here used only for purposes of indirectly estimating the dominating measure, not for directly estimating the structural growth functions.

4.2 Estimation of future risk given past events.

Fix $g \geq 1$ and $0 \leq s < g$. The marginal likelihood of the event

$[I=g, A^{(g)}, S_g=s]$ is

$$c_g(s, A^{(g)}) \cdot \int_{-\infty}^{\infty} \frac{\exp(\lambda s)\, dH(\lambda)}{\prod_{i=1}^{g} \{1 + \exp[\lambda + \phi_{\beta(i)}(i, A_i)]\}}$$

where $c_g(s, A^{(g)})$ is defined at (2). The fundamental general empirical Bayes identity is

$$(5)\quad \frac{P[S_g = s+1 \mid I = g, A^{(g)}]/c_g(s+1, A^{(g)})}{P[S_g = s \mid I = g, A^{(g)}]/c_g(s, A^{(g)})} = \frac{\int \frac{\exp[\lambda(s+1)]\, dH(\lambda)}{\prod \{1 + \exp[\lambda + \phi_{\beta(i)}(i, A_i)]\}}}{\int \frac{\exp(\lambda s)\, dH(\lambda)}{\prod \{1 + \exp[\lambda + \phi_{\beta(i)}(i, A_i)]\}}}$$

$$= E[e^{\lambda} \mid I=g, A^{(g)}, S_g=s] = \theta_s(g, A^{(g)}).$$

Thus the expected odds given number of term births, abortions, and age can be consistently estimated in large samples given the weights $c_g(s, A^{(g)})$ and consistent estimates of the $A^{(g)}$-specific point probabilities for S_g. This is considered below. The posterior expected odds on miscarriage at future pregnancy i and age A under assumption A4 is

$$(6)\quad E[\exp(\lambda + \phi_{\beta(i)}(i,A)) \mid I=g, A^{(g)}, S_g = s] = \theta_s(g, A^{(g)}) \cdot \exp[\phi_{\beta(i)}(i,A)].$$

Given an estimate $\hat{\theta}_g(g, A^{(g)})$ of the posterior expected odds on miscarriage at pregnancy 1, age A_0, we may estimate (6) by $\hat{\theta}_g(g, A^{(g)}) \cdot \exp[\hat{\phi}_{\beta(i)}(i,A)]$.

We view the outcome $S_g = s$ as a multinomial response taking on one of g+1 possible values s=0,1,...,g with respect to the dominating measure placing mass $c_g(s, A^{(g)})$ at the point s. Using sufficient statistics

$$t_j(s) = \begin{cases} 1 & \text{if } s \geq j \\ 0 & \text{if } s < j \end{cases} \qquad (j=1,\ldots,g)$$

the natural parameters are the quantities $\zeta_s(g,A^{(g)}) = \log \theta_s(g,A^{(g)})$ for $s = 0,\ldots,g-1$. These may be estimated by maximum likelihood in a multiple logistic regression model for the multinomial response. The exact specification of the age dependence in $\zeta_s(g,A^{(g)})$ is not known as it depends on the unknown prior H. However it seems reasonable for empirical work to use a Taylor approximation of the form

$$\zeta_s(g,A^{(g)}) \quad \alpha_s(g) + \sum_{i=1}^{g} \gamma_{is}(g)\,(A_i - A_0).$$

Further refinements are possible such as isotonizing the parameters $\zeta_s(g,A^{(g)})$ in s for purposes of reducing the mean squared error of the estimates.

Acknowledgement. I wish to thank Dr. Jennie Kline for her insightful epidemiolgic contributions and patient guidance during the preparation of this work.

REFERENCES

Cressie, N. (1982) A useful empirical Bayes identity. Ann. Statist. **10** 625-629.

Hui, S.L. and Berger, J. (1983). Empirical Bayes estimation of rates in longitudinal studies. J. Amer. Statist. Assoc. **78** 753-760.

James, W.H. (1969). Testing for birth-order effects in the presence of birth limitation or reproductive compensation. Appl. Statist. **18** 276-281.

James, W.H. (1974). Spontaneous abortion and birth order. J. Biosocial Sci. **6** 23-41.

Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., and Lehman, R.G. (1977). A general methodology for the analysis of repeated measurement of categorical data. Biometrics **33** 133-158.

Korn, E.L. and Whittemore, A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. Biometrics **35** 795-804.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. Biometrics **38** 963-974.

Leridon, H. (1976). Facts and artifacts in the study of intra-uterine mortality: a reconsideration from pregnancy histories. Population Studies **30** 319-335.

Naylor, A.F. (1974). Sequential aspects of spontaneous abortion: maternal age, parity, and pregnancy compensation artifact. Social Biology **21** 195-204.

Robbins, H. (1956). An empirical Bayes approach to statistics. Proc. Third Berkeley Symp. Math. Statist. Probab. **1** 157-163. Univ. of Calif. Press.

Robbins, H. (1980). An empirical Bayes estimation problem. Proc. Natl. Acad. Sci. U.S.A. **77** 6988-6989.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. Ann. Statist. **11** 713-723.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random effects models for serial observations with binary response. Biometrics **40** 961-971.

Warburton, D. and Fraser, F.C. (1964). Spontaneous abortion risks in man: data from reproductive histories collected in a medical genetics unit. Human Genetics **16** 1-23.

Wilcox, A.J. and Gladen, B.C. (1982). Spontaneous abortion: the role of heterogeneous risk and selective fertility. Early Human Devel. **7** 165-178.